Journal of Digital Imaging (2021) 34:1376–1386
https://doi.org/10.1007/s10278-021-00523-5

ORIGINAL PAPER

Rapid Quality Assessment of Nonrigid Image Registration Based on Supervised Learning

Eung-Joo Lee[1] · William Plishker[2] · Nobuhiko Hata[3] · Paul B. Shyn[3] · Stuart G. Silverman[3] · Shuvra S. Bhattacharyya[1,4] · Raj Shekhar[2,5]

Received: 21 October 2020 / Revised: 3 August 2021 / Accepted: 17 August 2021 / Published online: 13 October 2021

Abstract

When preprocedural images are overlaid on intraprocedural images, interventional procedures benefit in that more structures are revealed in intraprocedural imaging. However, image artifacts, respiratory motion, and challenging scenarios could limit the accuracy of multimodality image registration necessary before image overlay. Ensuring the accuracy of registration during interventional procedures is therefore critically important. The goal of this study was to develop a novel framework that has the ability to assess the quality (i.e., accuracy) of nonrigid multimodality image registration accurately in near real time. We constructed a solution using registration quality metrics that can be computed rapidly and combined to form a single binary assessment of image registration quality as either successful or poor. Based on expert-generated quality metrics as ground truth, we used a supervised learning method to train and test this system on existing clinical data. Using the trained quality classifier, the proposed framework identified successful image registration cases with an accuracy of 81.5%. The current implementation produced the classification result in 5.5 s, fast enough for typical interventional radiology procedures. Using supervised learning, we have shown that the described framework could enable a clinician to obtain confirmation or caution of registration results during clinical procedures.

Keywords Multimodality image registration · Quality assessment · Registration quality metric · Supervised learning

Introduction

Computed tomography (CT) is the imaging modality of choice for guiding percutaneous tumor ablation of liver tumors [1–3]. CT is common due to its ability to provide fast, high-resolution, and three-dimensional (3D) images of organs of interest intraprocedurally. In addition, CT images facilitate a radiologist's spatial understanding of the tumor inside the host organ and with respect to surrounding structures [4]. However, typically, only unenhanced CT images are used, so tumor margins may not be delineated well [5], which may contribute to misdirected or incomplete ablation [6–10].

Incomplete ablations or ablations with small margins, the shortest distance between the outer boundary of the tumor and the outer boundary of the ablation, have been correlated with high local tumor progression (LTP) rates [11, 12]. The effect of increasing the ablation margin on LTP can be dramatic. A study demonstrated that for hepatocellular carcinoma, an ablation margin increase from 1 mm to 3 mm corresponded to a drop in the LTP rate from 23% to 0% [13, 14]. Visualizing tumor boundaries clearly is important for achieving complete treatment with adequate ablation margins and thus best patient outcomes.

To delineate tumor boundaries intraprocedurally, we perform multimodality registration of preprocedural images such as magnetic resonance (MR) images to intraprocedural CT images. If successfully registered, tumor boundaries

✉ Eung-Joo Lee
elee1021@terpmail.umd.edu

Raj Shekhar
rshekhar@childrensnational.org

[1] Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA

[2] Institute for Advanced Computer Studies, University of Maryland, College Park, MD, USA

[3] IGI Technologies, Silver Spring, MD, USA

[4] Department of Radiology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

[5] Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Hospital, Washington, DC, USA

depicted on MR images and overlaid on live CT image can help an interventional radiologist guide the ablation needle precisely to the tumor [15]. Furthermore, the tumor boundaries can be directly compared to ablation effects depicted on intraprocedural CT images. The proposed registration can thus depict both tumor and ablation volumes [16]. Nonrigid image registration techniques can be used to correct accurately any misalignment between structures in the two images caused by physiologic variations [17]. For example, the liver may be misaligned because of diaphragmatic motion or different patient orientation (e.g., prone versus supine).

While nonrigid registration is a solution, registration accuracy, in some cases, may be affected by a number of external factors: image artifacts, major changes to anatomy, and improper initialization. To maintain clinically acceptable accuracy when such effects are unavoidable, a quality assessment system can be used to intercept poorly registered images and remove them from consideration by a clinician. The benefits of quality assessment for registration have been utilized in domains such as radiotherapy [18] and diagnostic radiology [19–21], but not yet in interventional radiology. In this work, we construct a solution based on accuracy metrics used in image registration that can be computed in near real time and combined to form a single assessment of multimodality registration quality as a binary value: *successful* or *poor*. We define a *successful* registration as one that is in close agreement with an expert-based correspondence and a *poor* registration as one that is not. Using expert-generated offline metrics of image registration, we present a supervised learning method, which predicts the quality of image registration from the trained classification model, and then test this framework on existing clinical data.

Our contribution is a quality assessment framework that is designed to integrate into existing interventional radiology workflows at interactive speeds and deliver fusion results that are quantifiably accurate relative to expert-validated solutions. We describe an implementation of the key component for this framework: a near real-time quality assessment module using supervised learning. By establishing a quality threshold enforced by our system, the fusion of MR and CT images to guide percutaneous ablations can be more reliable and enable faster and more accurate procedures.

Background

Automated Medical Image Registration

Image registration algorithms seek to find correspondences between two images and correct for any misalignments based on those correspondences. The task of discovering correspondences can be accomplished in a variety of ways including matching of extracted landmarks to homologous landmarks, an image to an atlas, or voxel intensities to voxel intensities. For a given similarity function S, we define an ideal registration transformation $\hat{t}$ by:

$$\hat{t} = \arg\max_{t \in T} S\left(I_r, I_f, t\right), \tag{1}$$

where I_r is the reference image (sometimes called fixed), I_f is the floating image (sometimes called moving), $t : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is a transformation that may be applied to points in the space of I_r to find the corresponding value or feature of a point in I_f. In this work, we focus on I_f and I_r when $m = 3$. T is the set of possible transformations, and $\hat{t}$ is a transformation which optimizes the similarity S being used. Note that t and $\hat{t}$ are transformations of reference image points to floating image points, such that the registered floating image will cover all the points in the reference image. Image registration methods commonly employ regularization to limit the set of transformations T. For instance, correspondences that indicate a transformation that is not physically possible are excluded from T because they cannot represent a possible correspondence even if they optimize a similarity function. T can represent rigid registration, in which only linear transformations are used to map the floating image to the reference, or nonrigid registration, in which nonlinear transformations are used to represent local deformations inside the volume. Rigid transformations may be represented by a tuple of parameters, while nonrigid transformations are often represented by a deformation field such that each voxel in the floating image has a unique vector to describe how it maps to the reference image space.

For multimodality image registration the focus of this work, we utilize a volume subdivision algorithm previously reported by our team [22] and accelerated on a GPU. This multilevel algorithm begins by performing a rigid registration between the two volumes (i.e., volumetric images). After rigid registration, the reference volume is divided into eight smaller subvolumes by dividing along each of the three axes, and an independent rigid registration takes place between each of the subvolumes and the floating image. By repeating the division and registration processes, rigid-body transformations that best model the local deformation for each of the subvolumes are determined. When these transformations are smoothly interpolated, nonrigid registration is achieved. Normalized mutual information (NMI) [23] is the similarity measure for registrations performed at all levels, and the solutions from previous levels contribute to NMI at subsequent subvolumes to improve local stability while retaining speed. This nonrigid registration algorithm is inherently tailored to benefit from the GPU′s parallel computing capability.

Metrics of Registration Quality

Regardless of the registration approach used and its demonstrated accuracy and robustness, suboptimal results are always possible. The ability to assess the quality of a given registration result is of critical importance in real-world applications. When evaluating registration approaches for accuracy, ideally, ground truth would be available. However, a true ground truth is difficult to be obtained unless the images or the transformation to be recovered is synthetic. For clinical images, proxies for ground truth fall into either offline or online metrics, which are described next.

Offline vs. Online Metrics

Relying on experts to mark points or contour structures is the gold standard of assessing registration quality when working with clinical images. Since an expert must create or at least approve marks, these quality metrics can only be used in an offline scenario, well after the registration is complete. Offline quality metrics based on such data cannot be fully automated and deployed with a registration solution. For such a situation, online metrics are required to evaluate the quality of a given resulting transformation. Assessments of registration quality are determined based on derived features of the image pair, evaluating correspondences or properties of the transformation itself. Online quality assurance metrics act as a guide to an automated optimization system toward a better solution contributing to either a similarity cost function or a regularization approach that augments such a function.

Point-Based Metrics

Target registration error (TRE) is the most common offline metric, in which an expert identifies point landmarks common to both images. The TRE is then the mean distance between matching point pairs, and it assesses the transformation accuracy between the reference and floating images at these point landmarks. The primary issue with using TRE is that point landmarks can be ambiguous. The dome of the liver may be identifiable in both images, but the soft-tissue deformation may cause a different physical point to be the dome. This can lead to difficulties identifying a true correspondence.

Contour-Based Metrics

Volumetric misalignment of an organ between two images can be evaluated using contour-based offline metrics such as Hausdorff distance (HD) and the Dice similarity coefficient (DSC) [24, 25]. The HD provides the mismatch distance between two contours, and perfect alignment yields an HD equal to 0. The DSC is defined as

$$DSC(A,B) = \frac{2|A \cap B|}{|A| + |B|}, \tag{2}$$

where A and B are labeled subvolumes in the same image space, $|A|$, $|B|$, and $|A \cap B|$ represent the volumes of the organ of interest in each image and the overlap between the two, respectively. Perfect alignment of the two data sets leads to a DSC value of 1. The HD and DSC are both derived from the same contours. However, they have distinctive interpretations; the HD indicates a misalignment of the region between the two images, whereas the DSC denotes what percent of the area in the registered image represents the true area. Although the HD provides an intuition as to the degree of misalignment in units of distance, it tends to be sensitive to outliers, which often makes it less stable than the DSC.

Intensity-Based Metrics

Intensity-based metrics are online metrics that either assume or discover correspondences between pixel intensities. They assess the similarity of an image pair for a given transformation. When image pixel correspondences (and hence intensity correspondences) are known, metrics such as the sum of differences or squared differences are effective at indicating similarity. When correspondences must be discovered, entropy-based features such as the mutual information (MI) can be employed, which can account for nonlinear dependencies in multimodality medical images [26]. In addition, normalized MI or NMI is effective for multimodality images that are sensitive to the changes in the overlap of low contrast regions [23]. The NMI is defined as

$$NMI(A,B) = \frac{H(A) + H(B)}{H(A,B)}, \tag{3}$$

where A and B indicate individual images, and H is a measure of individual or joint entropy. Although NMI has been proved to be effective, homogeneous regions in organs can be arbitrarily deformed as they lack features to anchor the correspondences.

Deformation Field-Based Metrics

For nonrigid registration, a smooth deformation field may be created from a transformation result. This field has properties that can be used to assess the quality of registration. When both a forward and a backward registration are performed, applying them in sequence should reproduce the original image albeit with few errors [27], which we refer to as self-TRE (STRE):

$$STRE(x_r) = |x_r - t_{backward}(t_{forward}(x_r))|, \tag{4}$$

where x_r is a point in the reference image, and $t_forward$ and $t_backward$ transform the point from the reference image space to floating image space, and vice versa. When sampled over an entire result, STRE provides a meaningful check of the consistency of the registration solution. Recovering the underlying deformation from two different directions and achieving a unique solution suggests that the solution is correct. If forward and backward registrations do not produce a self-consistent result that finds the same correspondences in both directions, then at least one of the registration results is incorrect. It is possible that the forward registration correspondence is correct but the backward registration result is incorrect, or vice versa. Even when the forward and backward registrations agree, it is also plausible that both are incorrect in the same way, overlooking an erroneous result that happens to be consistent. STRE is thus a necessary but not sufficient metric of accurate image registration.

The Jacobian determinant (JD) can be used to assess the amount of warping of a deformation field. In the context of image registration, the JD is defined as the determinant of the first-order partial derivative of the deformation field. In the case of image registration:

$$\mathrm{JD} = \det\left(\mathrm{J}_{ij}\right) = \det\left(\frac{\partial t_i}{\partial x_j}\right), \tag{5}$$

where $\mathrm{J}_i j$ are the entries in the Jacobian matrix, t_i is the value of the i component of the vector of the deformation field, and x_j is the j component of the point. When JD is applied on subvolumes of the deformation field, it indicates whether the deformation field is growing or shrinking. Variance in the size of subvolumes is possible, but for homogeneous regions that are erroneously deformed, the average, maximum, or minimum JD will correlate to such an error.

Related Work

Previous work on quality assurance of image registration has used Bayesian and supervised learning-based approaches [28]. During interventional procedures, a supervised learning-based assessment method is more suitable because of its lower computational complexity allowing the potential for real-time processing. In this section, we therefore review relevant previous studies on registration quality assessment using supervised learning. We then introduce our earlier registration study, the data, and the results of which were used in this work.

Previous Work on Quality Assessment

Wu and Samant [18] proposed a supervised learning-based approach to assess the quality of registration and identifying misregistration in patient positioning during radiation therapy. They used MI as a feature and an adaptive pattern classifier for quality assessment. Wu and Murphy [19] refined their previously reported work by adding more features and constructing a two-layer feed-forward neural network. With this framework, they focused on improving the evaluation of the quality of rigid registration of volumetric CT images for patient setup in radiotherapy. Shams et al. [21] presented a method for extracting a number of features in ultrasound images and used them to evaluate the quality of rigid registration for patient positioning during ultrasound-guided radiotherapy.

Heinrich et al. [29] and Muenzing et al. [20] presented a study on using statistical image features at distinctive landmark points of lung CT images and employing a set of different classifiers for registration quality assessment. They used manual landmark correspondences and evaluated the accuracy of spatial mapping between point landmarks in the CT images. Sokooti et al. [30] proposed a quality assessment framework using random regression forests. Using the regression model, they measured the registration error of chest CT scans in a quantitative manner, and then classified the registration quality. Schlachter et al. [31] presented a system for visualizing the quality of nonrigid registration of lung images. They used dissimilarity measures of local image patches and GPU acceleration to visualize the registration quality. Kybic et al. [32] proposed a prediction method that applies bootstrap resampling for image registration. Heinrich et al. [29] estimated registration uncertainty using supervoxel belief propagation to improve the accuracy of nonrigid registration.

Each of these works was motivated by the need for assessing whether a registration result is suitable for clinical use. However, these studies included only single-modality registration. These approaches are not easily applicable and extensible to multimodality registration, the focus of our work presented. Furthermore, there is no discussion of the computational aspects of these methods, and it is not clear if they run at practical speeds. In contrast, our work is focused on assessing the quality of multimodality registration results in near real time for interventional radiology procedures. This motivation necessitates a number of differences in the underlying methods for training and implementation.

Our Previous Registration Study

For the present work, we used imaging data from a previously published registration study [33] by our team. This study included ground-truth registration data provided by expert clinicians, board-certified interventional radiologists with over 10 years of experience each. The study focused on demonstrating the accuracy and speed of the

GPU-accelerated volume subdivision registration algorithm in the interventional radiology setting. We utilized abdominal imaging data that included a pair of volumetric intraprocedural CT and preprocedural MR images for each patient. The study was institutional review board-approved (Protocol 2002-P-001166/24), and the inclusion criteria were subjects who 1) had undergone CT-guided liver ablations between January 2013 and October 2013 and 2) had preprocedural MRI studies. Using these criteria, 14 subjects (aged 45–84 years; six men and seven women; one man underwent two ablations during two separate procedures) were included in the study. Tumor ablations were conducted using microwave ablation (n = 8; AMICA; HS Medical Inc., Boca Raton, FL), cryoablation (n = 6; Galil Medical Ltd., Yokneam, Israel), or radiofrequency ablation (n = 1; Covidien, Mansfield, MA). In one patient, both cryoablation and microwave ablation were performed to treat separate tumors in a single session. Using these datasets, automated multimodality registrations were performed between the MR and CT images, and then alignment was assessed against liver contours provided by clinical experts. This study showed that automatic registration provided by the GPU-accelerated volume subdivision algorithm was faster than semi-manual methods with no loss in accuracy.

For the purpose of quality assessment using supervised learning in the present study, datasets that lead to varying accuracies of multimodality registration need to be generated. To achieve this, we adopted the standard data augmentation strategy, and it is described in "Datasets".

Methods

We performed a retrospective study that included multimodality registration using existing data as discussed in "Our Previous Registration Study".

Real-Time Quality Assessment Architecture

To achieve the goal of real-time discrimination between successful and poor registration instances, we propose a novel quality assessment framework, as shown in Fig. 1. The key challenge in developing this framework is to compute and combine registration quality metrics rapidly with no manual assistance such that the metrics can be incorporated in a clinical workflow. Registration and fusion of MR and CT images that pass this framework should be of a quality similar to what an expert would have validated if given the time.

To meet this computational challenge, we use two GPUs concurrently to perform forward and backward registrations between the preprocedural MR image and the intraprocedural CT image. This approach enables near real-time computation of intensity-based and deformation field-based online metrics. Using these online quality metrics, we employ a supervised binary classifier based on the Random Forest method [34] for quality assessment. Random Forest is an ensemble learning method for classification, which constructs multiple decision trees and takes the average prediction over all of the trees to derive the classification result. For our framework, we selected Random Forests, as it does not require preprocessing such as data rescaling and feature selection of quality metrics. The proposed framework constructs a classifier which produces a binary assessment—successful or poor—of registration quality. This approach requires labeled datasets to train the classifier, which is explained next.

Experimental Setup for the Binary Classifier

Datasets

Positive and negative examples are needed to train the proposed binary classification model. In this work, positive and negative represent *successful registration* (SR) and *poor*

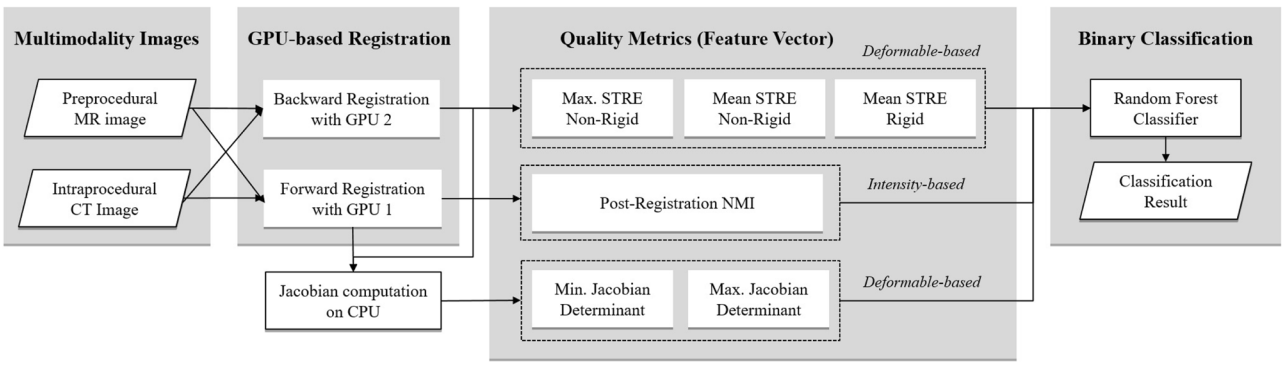

Fig. 1 Real-time quality assurance architecture using supervised learning. GPU-accelerated nonrigid registration was performed to compute registration quality metrics. Based on the quality metrics computed in real time, the binary classifier is constructed to evaluate the quality of multimodality registration results

registration (PR), respectively. We used expert-traced liver contours on preprocedural MR images and intraprocedural CT images to form SR and PR sets. We calculated the DSC based on liver contours for each registered MR image and its corresponding reference CT image. We divided the training data into two distinct sets using the following expert-derived offline metric:

$$\begin{aligned} \text{SR} &= \left\{t : DSC\left(V_r, V_{f_t}\right) \geqq th\right\}, \\ \text{PR} &= \left\{t : DSC\left(V_r, V_{f_t}\right) < th\right\}, \end{aligned} \tag{6}$$

where t is the transformation of the current registration, V_r is the labeled volume associated with the liver in the reference image, V_{f_t} is the labeled volume of the liver transformed from the floating image to the reference image, and th represents the threshold that separated SR and PR cases.

Following these criteria, we formed SR and PR examples through data augmentation from the clinical cases of our prior study. The overall approach was to perturb the initial misalignment or window-leveling of the MR and CT image pair, by perturbing corresponding parameters. The perturbation process, a form of data augmentation, enabled obtaining varying qualities of multimodality registration cases. Using empirical results, we found a range of each parameter that can be used to generate a mix of SR and PR cases. To window-level the MR and CT images, as is common, we rescaled their 16-bit intensity values to 8-bit intensity values using saturating logic (0 and 255) for voxel intensities outside of the window. We then obtained window-level candidates by randomly varying the high and low threshold values. This method allowed us to use window-level percentiles between 0.05 and 0.99 as lower and upper thresholds, respectively, which produced a mix of SR and PR cases. The slice size of MR and CT images was 512×512 pixels, and the number of slices varied between 20 to 45. MR pixel spacing was 0.70 mm with slice thickness of 3 – 5 mm. CT pixel spacing was 0.52 mm with slice thickness of 3 mm. The translation parameters were set from 2 mm to 3.5 mm along the three axes. Within these allowable ranges, we selected a window-level value and the offset of translation parameters at random assuming a uniform distribution. With this approach, prior to training, we set aside 200 image pairs generated from 2 cases for testing and generated 1256 image pairs for training from the remaining cases. Both sets comprised an equal number of SR and PR instances, and helped to construct and employ the classification model.

We defined a registration result as a successful if the DSC was above 0.84 and poor otherwise. **Quality threshold setting:** We derived the quality threshold of 0.84 from the results of the clinically acceptable algorithm as discussed in "Our Previous Registration Study". The previous study presented that the mean and standard deviation of the DSC were 0.89 and 0.05, respectively. The value 0.84 then is one standard deviation below the mean DSC, which defines a successful registration to be close to or about the average of an expert-derived solution. Figure 2 shows the distribution of DSC values for the entire datasets.

Figure 3 presents two cases of registration between preprocedural MR images and intraprocedural CT images. Each case shows an example of successful registration and two examples of poor registration resulting from perturbing the window-level and translation parameters. Figure 4 depicts the expert-generated liver contours of MR and CT images. We calculated the DSC value based on the overlay of the contours.

Supervised Learning Classifier

In this study, we calculated features as follows: post-registration NMI, STRE rigid registration, STRE nonrigid registration, and the Jacobian determinant. As shown in Fig. 1, we then obtained 6 online quality metrics using descriptive statistical features, which formed the input to train our Random Forest classifier. Using training datasets, we then constructed the binary classifier with k-fold cross-validation. We used the stratified fivefold cross-validation with a 4:1 ratio of training and validation sets. For Random Forest classification, we used 200 estimators with entropy as a splitting criterion and evaluated the classifier performance using a receiver operating characteristic (ROC) curve along with the computation of a confusion matrix. Based on the trained classifier, we also computed a confusion matrix using test datasets. From the confusion matrix of binary classification, we calculated sensitivity, specificity, and accuracy which represent, respectively, the rates of correctly classified SR, correctly classified PR,

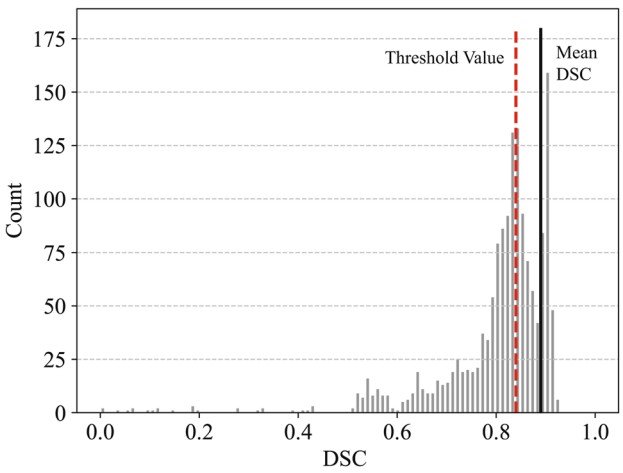

Fig. 2 Distribution of DSC values of the total training and test dataset. The quality threshold was defined as one standard deviation below (dashed line) the mean DSC value (solid line) acquired from the expert-derived solution [22]

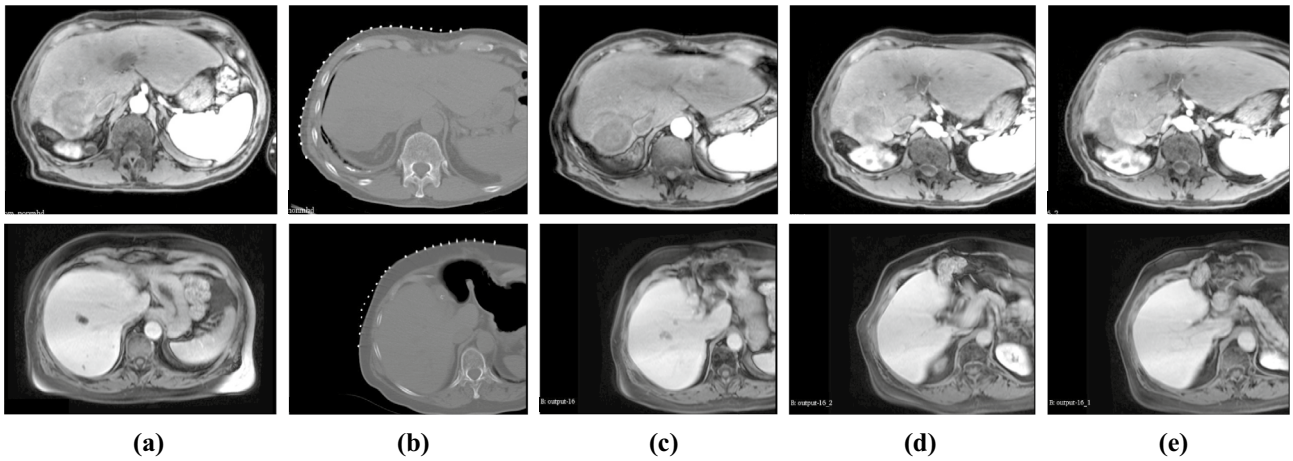

Fig. 3 Two examples of nonrigid registration of preprocedural MR and intraprocedural CT. **(a)** Preprocedural MR image, **(b)** Intraprocedural CT image, **(c)** Successfully registered MR image with a DSC of 0.89 (top) and 0.92 (bottom), **(d)** Poorly registered MR image with a DSC of 0.72 (top) and 0.78 (bottom), **(e)** Poorly registered MR image with a DSC of 0.68 (top) and 0.65 (bottom)

and correctly classified SR and PR cases. These measures demonstrated the performance of our classifier.

Platform

Forward and backward registrations were concurrently performed using a dual NVIDIA GTX 970 GPU and quad-core Intel Xeon 5140 CPU 2.33 GHz. We used the Insight Toolkit (ITK) [35] to calculate the Jacobian determinant and implemented the Random Forest classifier with k-fold cross-validation using the scikit-learn package (version 0.19) [36].

Regression Evaluation

Random Forest Regression

Using comparable metrics, we applied Random Forest regression to predict the DSC value for multimodality registration. A flowchart representation of the regression model is presented in Fig. 5. As Random Forest regression does not require feature selection and data rescaling, the quality metrics in the binary classifier can be applied without any pre- or post-processing. We created training and testing datasets by splitting the overall dataset with 50% of the samples

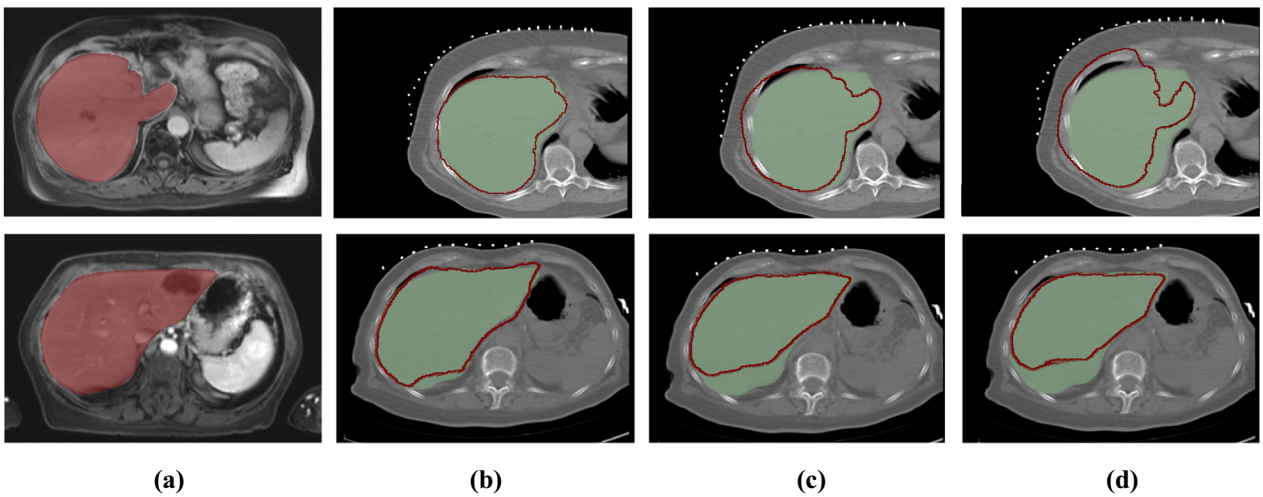

Fig. 4 Visualization of liver contours presented in the preprocedural MR image (red contours) and intraprocedural CT image (green region). **(a)** Preprocedural MR image with expert-defined liver region (red region), **(b)–(d)** Intraprocedural CT image, **(b)** Successful registration with a DSC of 0.89 (top) and 0.92 (bottom), **(c)** Poor registration with a DSC of 0.69 (top) and 0.78 (bottom), **(d)** Poor registration MR image with a DSC of 0.64 (top) and 0.65 (bottom). Note that the MR image is the floating image and the CT image is the reference image in these examples

selected randomly for training and the other 50% used for testing. We used the r-squared value for performance evaluation of the regression model.

Platform

The scikit-learn package (version 0.19) was used to implement the Random Forest regression model.

Results

Figure 6 shows the feature importance distribution of the registration quality metrics used in the Random Forest classifier. As can be seen, each feature contributes to the quality assessment of registration between the preprocedural MR image and the intraprocedural CT image.

Figure 7a has the ROC curve of the classifier. There are five separate ROC curves arising from fivefold cross-validation using training datasets, and the mean area under the ROC curves (AUC) is 0.94. Furthermore, Figure 7b illustrates the fitted linear regression, which identifies the relationship between the actual and predicted DSC values. We determined the associated r-squared value to be 0.89.

Tables 1 and 2 illustrate the confusion matrix for the assessment of registration quality from training and test datasets with sensitivity, specificity and accuracy.

Discussions

In this study, using online metrics that can be computed in real time, we developed a framework to assess the quality of nonrigid registration as either successful or poor. By using the expert-derived offline metrics as ground truth, we employed a Random Forest binary classifier for this assessment framework. We identified quality metrics that had the ability to differentiate between successful and poor registrations, and they formed the input to the framework. Furthermore, we ensured that the computation of each metric required no manual step and could be completed in a few seconds or less. These criteria led us to the selection of normalized mutual information, Jacobian determinant, and self-target registration error as the desired metrics. We calculated these quality metrics at different times during the registration process: before registration, between rigid and nonrigid stages, and after full nonrigid registration. We then used a machine learning method to combine these metrics so that each contributed to the overall classification.

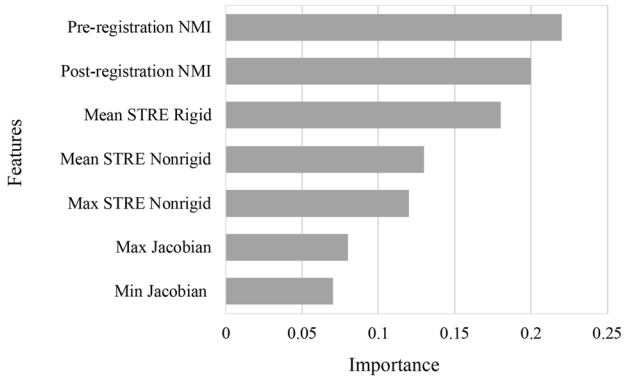

Fig. 6 Feature importance distribution of registration quality metrics

Each quality metric contributed to detecting registration failure modes. For instance, STRE (both rigid and nonrigid) detects misalignment between the MR and CT images post registration, whereas JD detects over-warping. The relatively even distribution of contributions observed in Fig. 6 indicates that each metric played a role in quality assessment. The most significant metric was found to be post-registration NMI, which was valuable in determining how likely it was that the registration had been successful. The lowest contributing factor in the classifier was JD. A possible explanation for this could be the use of DSC as the ground truth. The DSC metric is based on the matching of organ edges and/or surfaces, not the intra-organ deformation that may occur and that JD is sensitive to.

Based on these quality metrics, we constructed a binary classifier that can differentiate successful registrations from

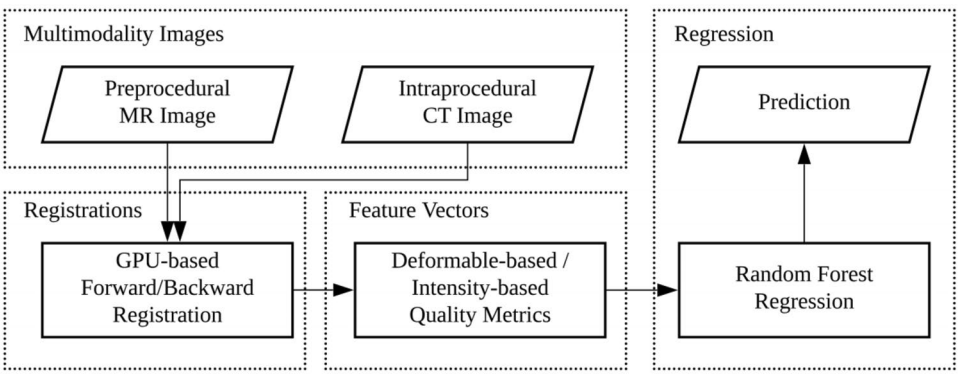

Fig. 5 Flowchart of the quality assessment process using Random Forest regression. The regression model is constructed from the metrics used for the binary classifier

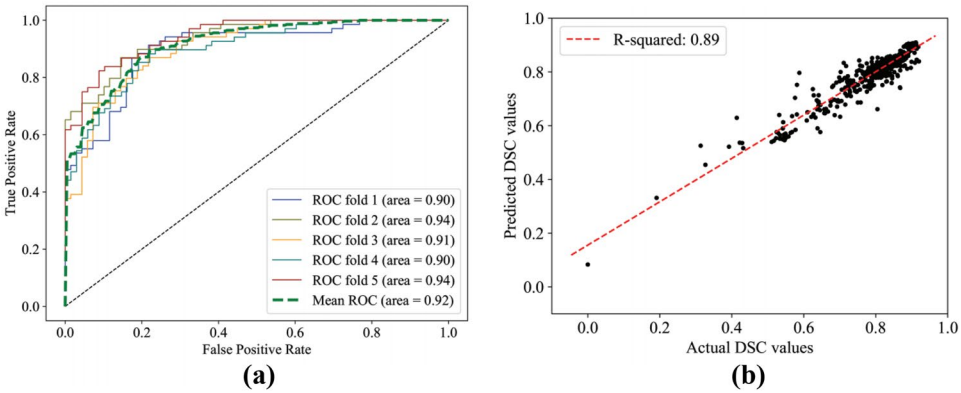

Fig. 7 **(a)** ROC curve of fivefold cross-validation for the binary classifier. The mean ROC curve is calculated by averaging validation curves of all of the folds. **(b)** Scatter plot of the actual and predicted DSC values from the Random Forest regression model. The result of linear regression is depicted by the dashed line

poor ones with a sensitivity of 88%, a specificity of 75%, and an accuracy of 81.5% (see Table 2). We examined the instances of false positives and false negatives in the test data. All instances of false negatives were from registration results that generated DSC values less than 10% away from our threshold of 0.84 DSC. On the other hand, all instances of false positives were less than 5% away from this threshold in terms of the actual DSC of the final registration. Even when false positives and false negatives occur, they appear close to the threshold set for distinguishing a successful registration from a poor registration result.

From our experimental results, we determined the r-squared value of our random-forest-based quality assessment system to be 0.89. This result indicates that the system is effective in predicting actual DSC values, and that the system can be also utilized to optimize registration using machine-learning-based regularization.

We have taken as an assumption that the DSC value is a good proxy for clinical acceptance of results. We have performed a small-scale test of this assumption with the help of a practicing radiologist. In our test, we created an equal mix of high quality and low quality registration cases, randomly selected from the evaluation pool. Using the corresponding fusion images, a board-certified radiologist rated them with the same lexicon. Of the 10 cases presented, the radiologist was confident about the labels for 7 of the cases; our registration algorithm agreed with all 7 of these cases. In the remaining 3 cases, the radiologist observed several features that indicated a successful registration, while other features led the radiologist to think it was a poor registration. These cases corresponded with 1 SR and 2 PR, each of which were close to our threshold, which leads us to the same conclusion that, while the threshold for acceptability may change per a specific clinician's preference, the basic ordering we have derived from DSC, we believe, is representative of registration quality.

In the results of our small-scale test, borderline cases according to our algorithm coincide with borderline cases for the clinician as well—that is, they correspond to the three cases that the radiologist was not confident about. How such borderline cases are ultimately treated in a clinical setting may require some tuning, and such a tuning-integrated approach fits within the framework presented in our paper. Larger-scale testing with clinicians as well as software tools to assist in tuning are useful directions for future work that build naturally upon the developments of this paper.

The results described above demonstrate that the proposed framework could effectively provide real-time confirmation or caution of registration results to the clinician during interventional procedures. Whereas the true positive and true negative cases are self-explanatory in this classification context, false negative and false positive cases need

Table 1 Confusion matrix of training datasets for the quality assessment of multimodality registration: successful registration (SR) and poor registration (PR). SR and PR correspond to positive and negative cases, respectively, for the confusion matrix

		Actual Class	
		SR (Positive)	PR (Negative)
Predicted Class	SR	83.6% (525/628)	14.0% (88/628)
	PR	16.4% (103/628)	82.8% (520/628)
Measure		Sensitivity = 83.6%	
		Specificity = 82.8%	
		Accuracy = 83.2%	

Table 2 Confusion matrix of test datasets for the quality assessment of multimodality registration: successful registration (SR) and poor registration (PR)

		Actual Class	
		SR (Positive)	PR (Negative)
Predicted Class	SR	88.0% (88/100)	12.0% (12/100)
	PR	25.0% (25/100)	75.0% (75/100)
Measure		Sensitivity = 88.0%	
		Specificity = 75.0%	
		Accuracy = 81.5%	

elaboration. A false negative means a valid fusion image will not be presented to the clinician, and/or the framework will improperly trigger another registration attempt. This implies the lost value of a valid fusion image being considered toward the clinical end. In this case, however, the registration engine can keep trying or reset when another intraprocedural image is taken. The harm of a false negative is principally the slowing down of a procedure. A false positive has a greater potential of harm, potentially misleading a clinician, who relies on the fusion image for a clinical decision. One approach to address this problem would be to bias the final system toward minimizing false positives. A larger validation dataset would enhance the sensitivity and specificity by mitigating over-fitting. The sensitivity and specificity metrics will inform clinicians as to the appropriate reliance on fusion as an adjunct guidance tool.

The computation time for the proposed framework is fast enough to be used clinically. It takes less than 5 seconds to compute the described metrics, each of which can be calculated concurrently, and an additional 0.5 seconds to obtain the classification result. The mean time to generate the outcome with our framework is 4.7 seconds. This mean value was derived by averaging over 200 image-pairs; the associated standard deviation is 0.27 seconds. The current implementation is sufficiently fast to provide on-demand automated quality assessment in a typical interventional radiology workflow. The execution of the framework can indeed be further accelerated with additional computing resources. When a preprocedural image is being registered with the latest navigational image during an interventional radiology procedure, the framework can be executed fast enough to allow successfully registered fusion images and prevent the display of poorly registered images. When poor registration is detected, the registration can be restarted with a different set of parameters or from a different starting position. As the computation time is further reduced, the framework has the potential to evolve into a machine learning-based regularization such that the proposed quality assessment can be incorporated directly into traditional registration algorithms.

We have focused on a single registration algorithm in this work. The novelty of our proposed method lies more on the framework of multimodality registration quality assessment and less on the registration algorithm itself. In addition, the presented framework is not reliant on any particular feature of the registration algorithm. Thus, it is generalizable to other registration algorithms. Indeed, an interesting direction for future work is experimentation with the proposed framework in conjunction with different registration algorithms.

For supervised learning, we augmented the training and test sets with synthetic samples generated from clinical imaging data from our previous study. This might have limited the overall performance of the classifier, as it might have over-fitted to the registration scenario and the ground truth metric. The proposed framework can be further extended and improved by using larger clinical datasets. A larger number of training samples may allow for exploring different classifiers and constructing deep neural networks, which may further augment our metrics and assess the registration quality in more anatomically general environment. In addition, a larger image database with clinical outcomes would enable clinical benefits to be part of the evaluation or training of our framework. Also, we used DSC as the ground truth for supervised learning. The framework can be further enhanced with additional quantitative distance-based metrics, such as the Hausdorff distance or center of mass distance, and landmark-based metrics, such as TRE. Using such metrics, we can expand the proposed framework with the validation of different registration algorithms. Moreover, new images being acquired and fused together with the clinical feedback being gathered will provide a logical next step in further development and evaluation of our proposed quality assessment framework.

Conclusion

In this paper, we presented a supervised learning-based framework that has the ability to assess the quality of multimodality nonrigid image registration results at interactive speeds. We proposed a framework that includes a Random Forest classifier constructed by using existing quality metrics that can be computed in real time, and then demonstrated that the overall accuracy and speed of the classifier are appropriate for practical clinical implementations. When introduced clinically, our platform would add a new level of confidence and sophistication to the use of image fusion in practice, enabling or improving on a myriad of image-based clinical applications.

Author Contributions In this work, we report on a novel image registration quality assessment framework designed to integrate into existing interventional radiology workflows and deliver image fusion results that are quantifiably accurate relative to expert-validated solutions. This framework was constructed based on image registration accuracy metrics that can be computed in near real time and combined to form the assessment of multimodality registration quality using supervised learning. This work is significant because it adds a critical quality control step in clinical implementation of multimodality image registration and fusion. By establishing a quality threshold enforced by our framework, the fusion of MR and CT images will be more reliable in our specific implementation and enable faster and more accurate procedures. The manuscript is entirely original and has not been copyrighted, published, or accepted for publication elsewhere.

Funding No funding was received with this work.

Data Availability Available upon request.

Code Availability Available upon request.

Declarations

Ethics Approval The study was institutional review board-approved (Protocol 2002-P-001166/24).

Consent to Participate Informed consent to participate in the study was obtained from participants.

Consent for Publication The patient, or parent, guardian or next of kin (in case of deceased patients) provided written informed consent for the publication of any associated data and accompanying images.

Conflicts of Interest/Competing Interests William Plishker and Raj Shekhar are founders of IGI Technologies, a medical technology start-up company. Other authors have nothing to disclose.

References

1. B. Furlow, Radiologic technology 90(6), 581CT (2019)
2. G. Antoch, H. Kuehl, F.M. Vogt, J.F. Debatin, J. Stattaus, Journal of vascular and interventional radiology: JVIR 13(11), 1155 (2002)
3. M. Sato, Y. Watanabe, K. Tokui, K. Kawachi, S. Sugata, J. Ikezoe, The American journal of gastroenterology 95(8), 2102 (2000). https://doi.org/10.1111/j.1572-0241.2000.02275.x
4. S. Goshima, M. Kanematsu, H. Kondo, R. Yokoyama, T. Miyoshi, H. Nishibori, H. Kato, H. Hoshi, M. Onozuka, N. Moriyama, AJR. American journal of roentgenology 187(1), W25 (2006). https://doi.org/10.2214/AJR.04.1878
5. D.M. Paushter, R.K. Zeman, M.L. Scheibler, P.L. Choyke, M.H. Jaffe, L.R. Clark, AJR. American journal of roentgenology 152(2), 267 (1989). https://doi.org/10.2214/ajr.152.2.267
6. M. Montorsi, R. Santambrogio, P. Bianchi, M. Donadon, E. Moroni, A. Spinelli, M. Costa, J Gastrointest Surg 9(1), 62 (2005). DOI S1091-255X(04)00452-4[pii]10.1016/j.gassur.2004.10.003
7. M.S.S. Chen, J.Q.Q. Li, Y. Zheng, R.P.P. Guo, H.H.H. Liang, Y.Q.Q. Zhang, X.J.J. Lin, W.Y. Lau, Ann Surg 243(3), 321 (2006). https://doi.org/10.1097/01.sla.0000201480.65519.b8
8. M. Abu-Hilal, J.N. Primrose, A. Casaril, M.J. McPhail, N.W. Pearce, N. Nicoli, J Gastrointest Surg 12(9), 1521 (2008). https://doi.org/10.1007/s11605-008-0553-4
9. H.H. Liang, M.S. Chen, Z.W. Peng, Y.J. Zhang, Y.Q. Zhang, J.Q. Li, W.Y. Lau, Ann Surg Oncol 15(12), 3484 (2008). https://doi.org/10.1245/s10434-008-0076-y
10. Z.W. Peng, Y.J. Zhang, M.S. Chen, X.J. Lin, H.H. Liang, M. Shi, Eur J Surg Oncol 36(11), 1054 (2010). DOI S0748-7983(10)00475-0[pii]10.1016/j.ejso.2010.08.133
11. Q. Yang, H. Qi, R. Zhang, C. Wan, Z. Song, L. Zhang, W. Fan, Journal of Vascular and Interventional Radiology 28(4), 481 (2017). https://doi.org/10.1016/j.jvir.2016.11.042
12. V.S. Sotirchos, L.M. Petrovic, M. Gönen, D.S. Klimstra, R.K.G. Do, E.N. Petre, A.R. Garcia, A. Barlas, J.P. Erinjeri, K.T. Brown, A.M. Covey, W. Alago, L.A. Brody, R.P. DeMatteo, N.E. Kemeny, S.B. Solomon, K.O. Manova-Todorova, C.T. Sofocleous, Radiology 280(3), 949 (2016). https://doi.org/10.1148/radiol.2016151005. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5006720/. 27010254[pmid]
13. Y.S. Kim, W.J. Lee, H. Rhim, H.K. Lim, D. Choi, J.Y. Lee, AJR Am J Roentgenol 195(3), 758 (2010). https://doi.org/10.2214/AJR.09.2954
14. G. Laimer, P. Schullian, N. Jaschke, D. Putzer, G. Eberle, A. Alzaga, B. Odisio, R. Bale, European radiology 30(5), 2463 (2020)
15. D. Wei, S. Ahmad, J. Huo, P. Huang, P.T. Yap, Z. Xue, J. Sun, W. Li, D. Shen, Q. Wang, Medical image analysis 65, 101763 (2020). https://doi.org/10.1016/j.media.2020.101763
16. H. Elhawary, S. Oguro, K. Tuncali, P.R. Morrison, S. Tatli, P.B. Shyn, S.G. Silverman, N. Hata, Acad Radiol 17(11), 1334 (2010). https://doi.org/10.1016/j.acra.2010.06.004
17. T. Rohlfing, C.R. Maurer Jr., W.G. O'Dell, J. Zhong, C.R. Maurer, W.G.O. Dell, Med Phys 31(3), 427 (2004)
18. J. Wu, S.S. Samant, Medical physics 34(6Part1), 2099 (2007)
19. J. Wu, M.J. Murphy, Medical physics 37(11), 5756 (2010)
20. S.E. Muenzing, B. van Ginneken, K. Murphy, J.P. Pluim, Medical image analysis 16(8), 1521 (2012)
21. R. Shams, Y. Xiao, F. H´ebert, M. Abramowitz, R. Brooks, H. Rivaz, IEEE transactions on medical imaging 37(2), 428 (2018)
22. V. Walimbe, R. Shekhar, Medical Image Analysis 10(6), 899 (2006). https://doi.org/10.1016/j.media.2006.09.002
23. C. Studholme, D.L.G. Hill, D.J. Hawkes, Pattern Recognition 32(1), 71 (1999). https://doi.org/10.1016/S0031-3203(98)00091-0
24. A. Bharatha, M. Hirose, N. Hata, S.K.Warfield, M. Ferrant, K.H. Zou, E. Suarez-Santana, J. Ruiz-Alzola, A. D'Amico, R.A. Cormack, R. Kikinis, F.A. Jolesz, C.M. Tempany, Med Phys 28(12), 2551 (2001)
25. K.H. Zou, S.K. Warfield, A. Bharatha, C.M. Tempany, M.R. Kaus, S.J. Haker, W.M. Wells 3rd, F.A. Jolesz, R. Kikinis, Acad Radiol 11(2), 178 (2004)
26. J.P. Pluim, J.A. Maintz, M.A. Viergever, IEEE transactions on medical imaging 22(8), 986 (2003)
27. T. Gass, G. Székely, O. Goksel, Journal of Medical Imaging 2(1), 014005 (2015)
28. R.D. Datteri, Y. Liu, P.F. D'Haese, B.M. Dawant, IEEE transactions on medical imaging 34(1), 86 (2015)
29. M.P. Heinrich, I.J. Simpson, B.W. Papież, M. Brady, J.A. Schnabel, Medical image analysis 27, 57 (2016)
30. H. Sokooti, G. Saygili, B. Glocker, B.P. Lelieveldt, M. Staring, in International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer, 2016), pp. 107–115
31. M. Schlachter, T. Fechter, M. Jurisic, T. Schimek-Jasch, O. Oehlke, S. Adebahr,W. Birkfellner, U. Nestle, K. Bühler, IEEE transactions on medical imaging 35(10), 2319 (2016)
32. J. Kybic, IEEE Transactions on Image Processing 19(1), 64 (2010)
33. J. Tokuda, W. Plishker, M. Torabi, O.I. Olubiyi, G. Zaki, S. Tatli, S.G. Silverman, R. Shekher, N. Hata, Academic Radiology 22(6), 722 (2015). https://doi.org/10.1016/j.acra.2015.01.007. http://linkinghub.elsevier.com/retrieve/pii/S1076633215000434
34. L. Breiman, Mach. Learn. 45(1), 5 (2001). https://doi.org/10.1023/A:1010933404324
35. T. Yoo, M. Ackerman, W. Lorensen, W. Schroeder, V. Chalana, S. Aylward, D. Metaxas, R. Whitaker, in Studies in Health Technology and Informatics, vol. 85 (2002), vol. 85, pp. 586–592. https://doi.org/10.3233/978-1-60750-929-5-586
36. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Journal of Machine Learning Research 12, 2825 (2011)